Middle East Journal of Applied Science & Technology (MEJAST)
Volume 6, Issue 2, Pages 64-75, April-June 2023

Prediction of Road Accidents Using Machine Learning Algorithms

R. Vanitha[1] & M. Swedha[2*]

[1]*Assistant Professor,* [2]*UG Student,* [1,2]*Department of CSE, IFET College of Engineering, Villupuram, Tamilnadu, India.*
*Corresponding Author (M. Swedha) Email: swedha0025@gmail.com**

DOI: https://doi.org/10.46431/MEJAST.2023.6208

Article Received: 19 March 2023 *Article Accepted: 30 April 2023* *Article Published: 18 May 2023*

ABSTRACT

Today, one of the top concerns for governments is road safety. There are many safety features built into cars, yet traffic accidents still happen frequently and are unavoidable. To lessen the harm caused by traffic accidents, predicting their causes has become the primary goal. In this situation, it will be beneficial to examine the frequency of accidents so that we can use this information to further aid us in developing strategies to lessen them. From this, we can deduce the connections between traffic accidents, road conditions, and the impact of environmental factors on accident occurrence. In order to construct an accident prediction model, I used machine learning techniques, including the Decision Tree, Random Forest, and Logistic Regression. The development of safety measures and accident prediction will both benefit from these classification systems. Several elements, including weather, vehicle condition, road surface condition, and light condition, can be used to predict road accidents. Three dataset files—accidents, casualties, and vehicles are loaded into this dataset. This allows us to forecast the severity of accidents.

Keywords: Road accidents; Logistic regression; Factors; Machine learning; Random forest.

1. Introduction

The World Health Organization recently revealed fatality figures, and they show an alarming amount of traffic accidents worldwide every year. 1.2 million people were killed in automobile accidents. 50 million persons were hurt every year. Every day, there were roughly 3,300 fatalities and 137,000 injuries. 43 billion dollars in direct economic damages, as well as a direct threat to human life and property safety posed by traffic accidents that frequently occur.

One of the most crucial areas of traffic safety research is the prediction of road accidents. Road geometry, traffic flow, driver traits, and road environment all have a major impact on the likelihood of traffic accidents. Many research have been carried out to forecast accident frequencies and evaluate the elements of traffic accidents, including studies on the identification of dangerous locations/hot spots, the analysis of accident injury-severities, and the study of accident length. Several research concentrate on the accident's mechanism. Weather and the visibility of the road are other concerns.

India is experiencing an increase in accidents, which is very concerning. Recent data indicates that India is responsible for 6% of all traffic accidents worldwide. Two-wheelers' irresponsibility is a major cause of recorded accidents, and over-speeding is another contributing element. Accidents brought on by drunk driving or other types of traffic offences happen frequently as well. Despite having established rules and traffic codes, a number of accidents have been caused by people being careless with the speed of their vehicles, the condition of their vehicles, and their failure to wear helmets. Although the growing number of vehicles is the primary contributor to traffic accidents, the importance of the state of the roads and other environmental elements cannot be understated. In India, the amount of fatalities from traffic accidents is undoubtedly concerning. With more than 137,000 persons suffering injuries as a result of traffic accidents, the situation is exceedingly grim. This number is more than four times the number of people killed by terrorism each year. Accidents involving large trucks and even those

Middle East Journal of Applied Science & Technology (MEJAST)
Volume 6, Issue 2, Pages 64-75, April-June 2023

involving buses, which are utilised for public transit, are among the deadliest kinds of accidents that can happen and cost the lives of innocent people.

Rain, fog, and other weather conditions can greatly increase the likelihood of accidents. So, it would be easier to take action to reduce accidents if you have a proper estimation of incidents and are aware of accident hotspots and contributing variables. This necessitates a careful examination of events and the creation of models for predicting accidents. In this essay, we'll talk about how an accident prediction model can help determine the risks that go along with various scenarios involving accidents.

A probabilistic model linking important crash precursors to changes in crash potential was created by Lee et al. Using the matched case-control logistic regression technique, Abdel [10] created a previous crash prediction model. The traffic police have no unique method for predicting which location is accident-prone at a given time. In order to effectively plan and manage traffic, it is vital to predict traffic accidents because they often involve nonlinear factors like people, cars, roads, weather, and other factors that are very random. Because of the noise pollution and the small amount of data, typical linear analyses are unable to reflect the true situation, and the prediction's outcome is insufficient. Traditional BP networks suffer from flaws including local minimum, excessive iterations, delayed training, and other issues.

2. Related Works

Several researchers have employed deep learning for picture classification [1], text mining, fake news identification, and text classification, among other applications of data mining. Many scholars have thought about employing various data mining techniques to analyze traffic accident data. Several studies examined the severity of traffic accidents in various nations.

Using some parameter selection techniques, they chose the 16 parameters that have the greatest impact on the severity of the drivers' injuries out of the 150 total parameters. To categorise the degree of injury severity, ANN was used. They managed to obtain a somewhat poor accuracy of 40.71%. Regression models are the fundamental elements for any type of data analysis with the link between the explanatory variable and response variable. They discovered that the model's sensitivity and specificity were 40% and 98%, respectively, at the probability cut-point of 0.20. Also, their study's findings demonstrate the significance of velocity, seat belt use, and impact direction in determining how serious an accident would be. To analyse quality accidents, authors suggested fuzzy rule mining.

Three classes were forecasted as a collection of binary prediction models during the analysis process, helping to increase the projected model's accuracy, which was achieved at 60.94%. and the crucial factor that influences how serious accidents are is incorrect overtaking and seat belt neglect. Road accident analysis was done by Sharma et al. using a Support Vector Machine and multi-layer perception. They only used a few data samples for their experiments. They stated that the event was caused by high-speed driving while intoxicated. For classification and grouping, Tiwari et al. [11] employed machine learning models such as Decision Tree (DT), Naive Bayes (NB), and Support Vector Machine (SVM). With the cluster dataset, they got superior results.

The ability to predict the seriousness of incidents on the road is still being developed. Prediction accuracy will increase with the acquisition of a suitable strategy. Choosing the optimal paradigm also aids in determining the

causes of traffic accidents. Moreover, elements that are more pertinent to the aim can aid machine learning models in producing better predictions for outcomes that were not previously known.

3. Proposed System

A. Proposed Road Accident Prediction System

An Machine Learning approach that gauges accident severity from the circumstances. With 1.6 million accident records from 2005 to 2015, it has been trained. Accuracy improves with more information. Such a model's goal is to be able to foretell which circumstances will be more likely to result in accidents. In order to give quicker care and preventative measures, we will even aim to discover potential incidents with greater accuracy. The approaches for estimating the route from the provided dataset are discussed in this section. The classification algorithm, such as Logistic Regression, Decision Tree, or Random Forest, used to categorise a dataset for the prediction of traffic accidents. To anticipate the accident severity, I have tested three alternative algorithms. In terms of predicting all the classes of accident severity, it was evident that Decision Tree and Random Forest outperformed each other significantly. Although logistic regression is more accurate, this does not necessarily imply that it is the best algorithm. To forecast all the classes in the section on hyperparameter tuning, authors even attempted using multi-normal. Still, it only predicted one higher occurring class event. The conclusion is that Logistic regression has an accuracy of 86.23%, Decision tree has an accuracy of 75.26%, and Random forest has an accuracy of 86.86%. It is obvious that Random Forest, with its precise accuracy of 86.86%, produces the best results..picture enhancement are some of the pre-processing methods employed in this system.

The suggested system's block diagram is shown in Figure 1.

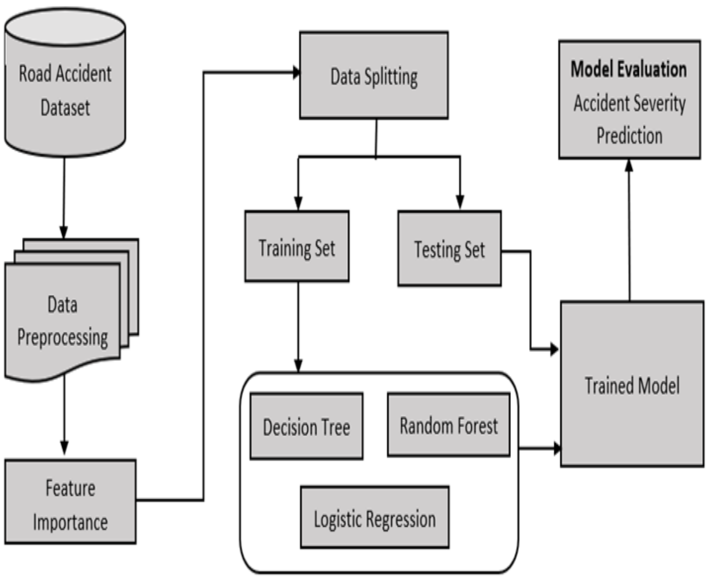


Figure 1. Block diagram of the proposed system

B. Data Importing

To analyse this data, three files are imported. This information is divided into three files: accidents, casualties, and vehicles. We do, however, have one more file that contains generic data concerning traffic counts from 2000 to 2015. For the machine learning portion, we can use data on general traffic patterns.

- The necessary package imports are completed.
- 3 CSV documents Accidents.csv Casualties.csv Vehicles.csv were used.
- Importing data into a data frame using pandas.

C. Applied Machine Learning Techniques

(i) Logistic Regression

The supervised classification technique known as logistic regression predicts a result that has two alternative values, such as zero or one, no or yes, false or true. The likelihood of a binary dependent variable being predicted from the dataset's independent variable is given by logistic regression. Despite the obvious similarities between logistic regression and linear regression, logistic regression produces a curve as opposed to a straight line. Using one or more independent variables or predictors, logistic regression creates logistic curves that represent values between 0 and 1. An analysis of the relationship between a number of independent factors and a categorical dependent variable is performed using a regression model called logistic regression.

$$p = \frac{e^{\alpha+\beta_n X}}{1 + e^{\alpha+\beta_n X}}$$

(ii) Decision Tree

A decision tree method uses conditional control statements to predict the final choice by creating a tree-like graph or model of options and possible outcomes. A decision tree is a tool for approaching discrete-valued target functions, and it is represented as a learned function. These algorithms are well recognised for facilitating inductive learning and have been successfully used to a number of tasks. The decision tree is evaluated against the transaction value before a path is displayed from the root node to the transaction's output or class label. This process is done for each new transaction to determine whether it is legitimate or fraudulent.

$$Entropy(S) = \sum_{i=1} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

(iii) Random Forest

A technique for classification and regression is called Random Forest. It is, in essence, a collection of decision tree classifiers. Random forest has an advantage over decision trees since it corrects the propensity of overfitting to their training set. After a decision tree has been built, each node is divided on a feature selected at random from the entire feature set. A subset of the training set is used to train each individual tree. Even for large data sets with various attributes and data occurrences, training is remarkably quick in a random forest since each tree is trained independently of the others. It has been found that the Random Forest approach is resistant to overfitting and provides a good approximation of the generalisation error.

$$\frac{1}{X} \sum_{x=1}^{x} f_x(\dot{R})$$

(iv) Hyper Parameter Tuning

The performance of ML algorithms for prediction can be significantly impacted by HP tuning [9]. A ML algorithm's HPs are often configured through a process of trial and error. Finding a decent set of numbers manually can take a lot of time, depending on how long the ML algorithm being used needs to train. Because of this, current research on HP for ML algorithms has focused on improving HP tuning strategies [5].

The HP process is typically viewed as an optimization (blackbox) issue, with the algorithm's objective function being the predictive accuracy of the resulting model.

D. Performance Evaluation Measures

(i) Accuracy

Several parameters have been represented in the evaluation of models, a crucial task in categorization. The most widely used evaluation criteria used in this study are accuracy, precision, recall, and F-score. Accuracy is defined formally as the correctness of a forecast and is computed as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy can be calculated in terms of positives and negatives in binary classification:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

(ii) Precision

The preciseness of a classifier is referred to as precision, and it indicates what proportion of all tuples with a positive label are genuinely positive. It is determined by:

$$\text{Precision} = \frac{TP}{TP + FP}$$

(iii) F-Measure/F1-Score

The F-score is a statistical analysis metric for categorization that computes a score between 0 and 1 while taking the classifier's recall and precision into account. It is calculated as follows to demonstrate the impact of both recall and precision:

$$F = \frac{2 \times \text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

(iv) Recall

On the other hand, recall is frequently referred to as the measure of completeness and it displays the proportion of true positive tuples that are correctly classified. It is determined by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

E. Data Pre-processing

(i) Data Cleaning

Here, we define noisy, pointless data. Visualization also helps us determine which elements are more crucial.

(ii) Data Visualization

The first thing we can do is learn the date and time of the collision, as well as the age of some of the drivers who were involved. We can determine the frequency of accidents based on the days of the week. Hours of the day can be used to determine the amount of accidents, and the driver's age can provide additional information.

(a) Accidents on Day of Week

The amount of accidents can be determined based on the days of the week. As we can see, from 2005 to 2015, Thursday had the largest number of accidents in this dataset. We must remember that the number of accidents may vary based on the volume of traffic on a certain day.

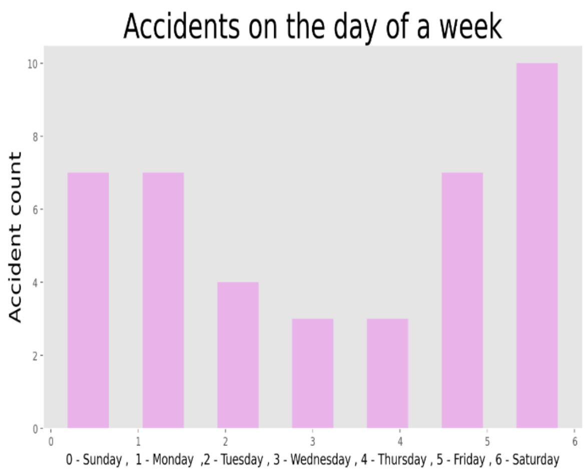


Figure 2. Accidents on the day of a Week

(b) Time of Accident

Here, we discovered that accidents tended to occur more frequently after midday. We can presume that the greatest traffic is flowing at this time of day because people are likely departing for work.

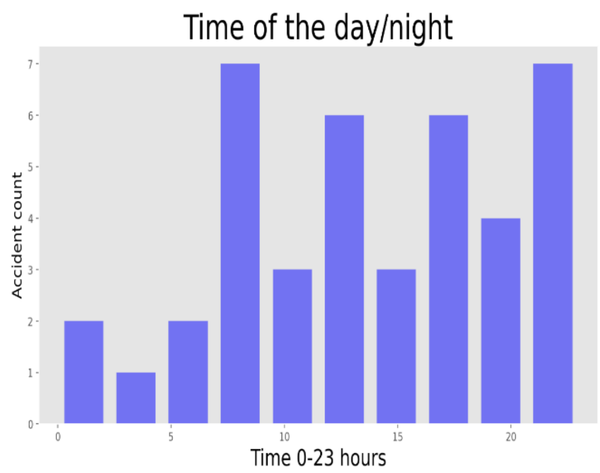


Figure 3. Time of Accident

(c) Age Band of Casualties

This fact regarding this dataset is quite intriguing. Most of the drivers who are in accidents are between the ages of 25 and 35. Nevertheless, we are unsure about the proportion of drivers between the ages of 25 and 35 compared to those of other ages. I would predict that there would be a higher proportion of drivers between the ages of 25 and 35.

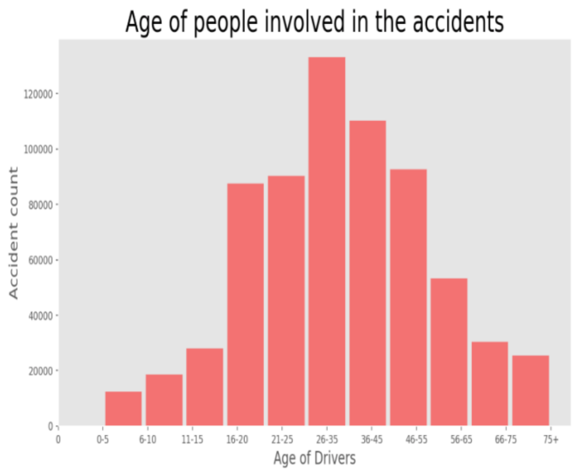


Figure 4. Age of People involved in accident

(d) Co-relation between variables

Our dataset only contains numeric values. We can determine whether two columns are correlated. We can see that there aren't many variables that have substantial relationships with one another. Speed restriction and Urban or Rural Area only have one substantial positive link.

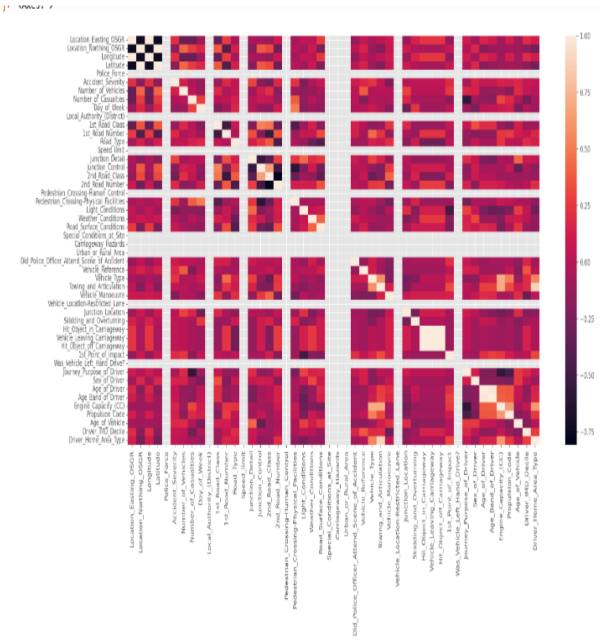


Figure 5. Co-relation

(e) Speed of Cars

The majority of collisions happened on roads with 30 mph or higher speed limits. Accidents may occur as a result of stop signs, lane changes, turning into parking lots, etc.

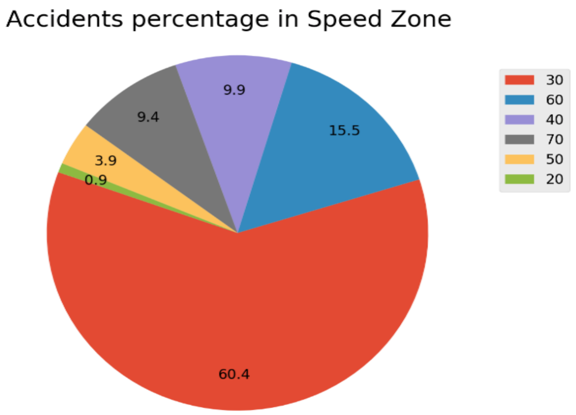
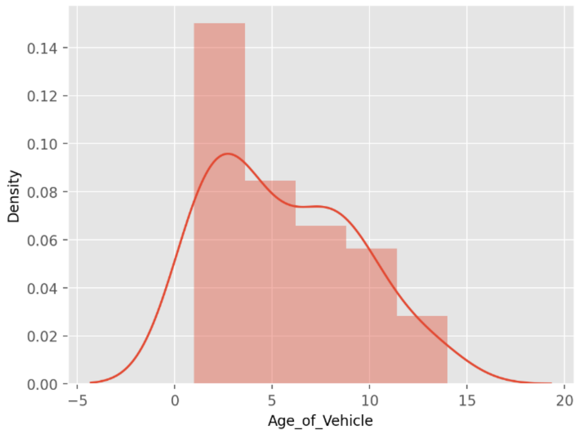


Figure 6. Accident Percentage

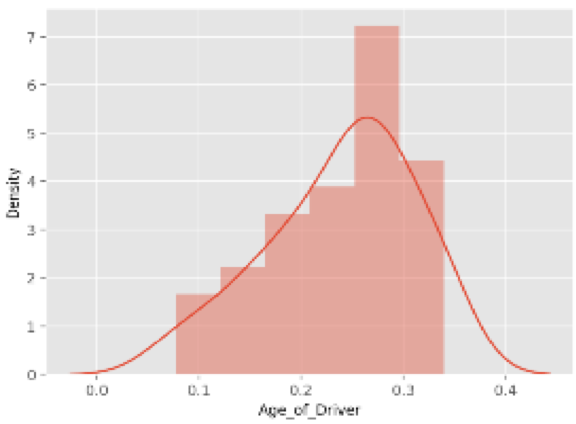
(f) Normalize the Data

We will standardise a small number of columns, so our machine learning algorithms won't be badly impacted. In the dataset, the age of the drivers ranges from 18 to 88, and we may standardise it. Also, the age of the car ranges from 0 to 100, which can affect how well your machine learning algorithm performs. We will normalise this prediction as well.

Before Normalization



After Normalization



4. Experimental Results

In order to conduct Python programming language experiments, I used Jupyter notebook environment. Sklearn is used to implement machine learning models. The outcomes of the Logistic Regression, Decision Tree, and Random Forest analyses performed on the dataset of traffic accidents are shown in this section.

(1) Logistic Regression Result

(Accuracy, 86.23)

	precision	recall	f1-score	support
1	0.000000	0.000000	0.000000	4111
2	0.000000	0.000000	0.000000	38151
3	0.862320	0.999996	0.926069	264697
micro avg	0.862317	0.862317	0.862317	306959
macro avg	0.287440	0.333332	0.308690	306959
Weighted avg	0.743596	0.862317	0.798568	306959

Predicted	**1**	**3**	**All**
Actual			
1	0	4111	4111
2	0	38151	38151
3	1	264696	264696
All	1	306958	306959

(2) Decision Tree Result

(Accuracy, 75.26)

	precision	recall	f1-score	support
1	0.036793	0.046217	0.040970	4111
2	0.158974	0.287780	0.172180	38151
3	0.871137	0.844921	0.857829	264697
micro avg	0.752550	0.752550	0.752550	306959
macro avg	0.355635	0.359639	0.356993	306959
Weighted avg	0.771451	0.752550	0.761672	306959

Predicted	**1**	**2**	**3**	**All**
Actual				
1	190	894	3027	4111
2	931	7164	30056	38151
3	4043	37006	223648	264696
All	5164	45064	256731	306959

(3) Random Forest Result

(Accuracy, 86.86)

	precision	recall	f1-score	support
1	0.031496	0.002928	0.005358	1366
2	0.195615	0.040622	0.067291	38151
3	0.884926	0.979143	0.929653	166321
micro avg	0.868601	0.868601	0.868601	188464
macro avg	0.370779	0.340898	0.356993	188464
Weighted avg	0.802781	0.868601	0.827884	188464

done

(4) Decision Tree Hyperparameter Tuning Result

```
(Accuracy, 85.71)
              precision    recall  f1-score   support

           1   0.071429  0.000730  0.001445      4111
           2   0.323387  0.045451  0.079700     38151
           3   0.866655  0.987333  0.923066    264697

   micro avg   0.857056  0.857056  0.857056    306959
   macro avg   0.420490  0.344504  0.334737    306959
Weighted avg   0.788483  0.857056  0.805904    306959

Predicted  1       2        3      All

   Actual
     1       3    301     3807    4111
     2      13   1734    36404   38151
     3      26   3327   261344  264697
     All    42   5362   301555  306959
```

(5) Logistic Regression Hyperparameter Tuning Result

```
(Accuracy, 86.23)
              precision    recall  f1-score   support

           1   0.000000  0.000000  0.000000      4111
           2   0.000000  0.000000  0.000000     38151
           3   0.862317  0.999974  0.926058    264697

   micro avg   0.862298  0.862298  0.862298    306959
   macro avg   0.287439  0.333325  0.308686    306959
Weighted avg   0.743594  0.862298  0.798558    306959

Predicted  1       3      All

   Actual
     1     0    4111    4111
     2     0   38151   38151
     3     7  264690  264697
     All   7  306952  306959
```

(6) Overall Accuracy

S. No.	Algorithm Name	Accuracy Score (%)
1.	Decision Tree	75.26
2.	Random Forest	86.86
3.	Logistic Regression	86.23
4.	Decision Tree Hyperparameter Tuning	85.74
5.	Logistic Regression Hyperparameter Tuning	86.23

5. Conclusion

The primary cause of injuries, fatalities, and property damage are traffic accidents, which have grown to be a serious problem for public health and safety. Accidents also contribute to traffic congestion and delays. It is necessary to manage accidents by looking at connected aspects in order to increase the effectiveness of the transportation system. In this study, a machine learning model called Logistic Regression, Random Forest, and Decision Tree is used to predict the severity level of traffic accidents. The experimental findings in this research

revealed why Random Forest classification results outperformed Logistic Regression and Decision Tree. The top features include distance, temperature, wind Chill, humidity, visibility, and wind direction, and significant features are identified using Random Forest and Decision Tree, respectively. Given that Random Forest consistently outperformed all ensemble models in predicting accident severity, it may be claimed that it is the most efficient and effective model of all. On the other hand, measuring the link between significant features and traffic accidents was the main goal of distinguishing significant features from general features. As a future work, more resources with continuous prediction and alerts can be sent to the police for every location at regular intervals of time to take preventive measures. It can be incorporated with Google Maps which can be live tracked by the police. A fully-fledged web app for user and police interaction can be published for use in real-time. It can be used for Indian states or cities, if proper data of accidents is provided by the Indian Government.

Declarations

Source of Funding

This study did not receive any grant from funding agencies in the public or not-for-profit sectors.

Competing Interests Statement

The authors have declared no competing interests.

Consent for Publication

The authors declare that they consented to the publication of this study.

Authors' Contribution

Both the authors took part in literature review, research, and manuscript writing equally.

References

[1] Y. Zou, B. Lin, X. Yang, L. Wu, M. M. Abid, and J. Tang (2021). Application of the Bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management. J. Adv. Transp., Pages 1–9.

[2] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. Anal. Methods Accident Res., 27.

[3] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G.S. Choi (2021). Predicting numeric ratings for Google apps using text features and ensemble learning. ETRI J., 43(1): 95–108.

[4] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G.S. Choi, and B.W. (2020). Fake news stance detection using deep learning architecture (CNNLSTM). IEEE Access, 8: 156695–156706.

[5] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi, and B.W. (2021). Aggression detection through deep neural model on Twitter. Future Gener. Comput. Syst., 114: 120–129.

[6] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G.S. Choi, and A. Mehmood (2020). Duplicate questions pair detection using Siamese MaLSTM. IEEE Access, 8: 21932–21942.

[7] M.I. Sameen and B. Pradhan (2017). Severity prediction of traffic accidents with recurrent neural networks. Appl. Sci., 7(6): 476.

[8] S. Seid and Pooja (2019). Road accident data analysis: Data preprocessing for better model building. J. Comput. Theor. Nanosci., 16(9): 4019–4027.

[9] S.K. Singh (2017). Road traffic accidents in India: Issues and challenges. Transp. Res. Proc., 25(5): 4708–4719.

[10] D. Delen, R. Sharda, and M. Bessonov (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Anal. Prevention, 38(3): 434–444.

[11] D.W. Kononen, C.A.C. Flannagan, and S.C. Wang (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Anal. Prevention, 43(1): 112–122.

[12] P. Duan, Z. He, Y. He, F. Liu, A. Zhang, and D. Zhou (2020). Root cause analysis approach based on reverse cascading decomposition in QFD and fuzzy weight ARM for quality accidents. Comput. Ind. Eng., 147.

[13] H.M. Alnami, I. Mahgoub, and H. Al-Najada (2021). Highway accident severity prediction for optimal resource allocation of emergency vehicles and personnel. In Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC), Pages 1231–1238.

[14] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. S. Choi (2021). Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. Comput. Intell., 37(1): 409–434.

[15] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi (2021). Discrepancy detection between actual user reviews and numeric ratings of Google app store using deep learning. Expert Syst. Appl., 181.

[16] P. Tiwari, S. Kumar, and D. Kalitin (2017). Road-user specific analysis of traffic accident using data mining techniques. In Proc. Int. Conf. Comput. Intell., Commun., Bus. Anal. New York, NY, USA, Pages 398–410.

[17] R.E. AlMamlook, K.M. Kwayu, M.R. Alkasisbeh, and A.A. Frefer (2019). Comparison of machine learning algorithms for predicting traffic accident severity. In Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol., Pages 272–276.

[18] T. Beshah and S. Hill (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. In Proc. AAAI Spring Symp., Artif. Intell. Develop., Volume 24, Princeton, NJ, USA: Citeseer, Pages 1173–1181.

[19] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang (2017). Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transp. Syst., 18(9): 2303–2310.

[20] B. Yu, Y.T. Wang, J.B. Yao, and J.Y. Wang (2016). A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. Neural Netw. World, 26(3): 271.